Auer, Tobias; Ulasik, Maria; Holzmeister, Felix

Working Paper

A Comment on "Motivated Errors" by Exley and Kessler (2024)

I4R Discussion Paper Series, No. 161

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Auer, Tobias; Ulasik, Maria; Holzmeister, Felix (2024) : A Comment on "Motivated Errors" by Exley and Kessler (2024), I4R Discussion Paper Series, No. 161, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/303193

INSTITUTE for REPLICATION

No. 161
I4R DISCUSSION PAPER SERIES

A Comment on "Motivated Errors" by Exley and Kessler (2024)

Tobias Auer

Maria Ulasik

Felix Holzmeister

September 2024

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 161

A Comment on "Motivated Errors" by Exley and Kessler (2024)

Tobias Auer[1], Maria Ulasik[1], Felix Holzmeister[1]

[1]University of Innsbruck/Austria

SEPTEMBER 2024

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

A comment on "Motivated Errors" by Exley and Kessler (2024)*

Tobias Auer
University of Innsbruck

Maria Ulasik
University of Innsbruck

Felix Holzmeister
University of Innsbruck

August 29, 2024

Abstract

This report evaluates the computational reproducibility and analytical robustness of Exley and Kessler's (2024) investigation into "motivated errors," which suggests that individuals may rationalize selfish behavior by attributing their errors to confusion. Using the original data and code, we could regenerate all results reported in the manuscript and online appendices with full precision. However, our re-analysis identified significant limitations, including insufficiently annotated code, ambiguous variable naming, and the absence of essential participant-level data, which obstruct comprehensive robustness checks. These challenges underscore the importance of best practices in data and code sharing to enhance the transparency and credibility of economic research. Our reflection not only contributes to discussions on empirical rigor but also advocates for improved standards in sharing scholarly resources.

Keywords: reproducibility, robustness, credibility, data/code sharing
JEL: C18, C81, C91, D91

* This comment was written during and after the *Innsbruck Replication Games* on July 5, 2024, organized by the Institute for Replication (www.i4replication.org). We thank Derek Mikola and Abel Brodeur for their assistance and helpful comments. The authors have no competing interests to declare. T.A. and M.U. contributed equally. Correspondence should be addressed to F.H. (felix.holzmeister@uibk.ac.at).

1. Introduction

A recent study by Exley and Kessler (2024a; E&K henceforth) investigates whether economic decision-makers appeal to the possibility of being confused or making an honest mistake to justify their behavior and benefit themselves. E&K provide evidence for "motivated errors" in three sets of experiments (carried out on *Amazon Mechanical Turk*): *(i)* In the *Adding* study, participants were asked to choose between receiving money for themselves and donating to a charity. E&K find that participants are less likely to select the charity when a zero is added to the donation amount. When selfish motives were removed by asking participants to choose between two donation amounts, participants' choices were not affected by the addition of zero. *(ii)* In the *Correlation Neglect* study, building upon the design in Enke and Zimmermann (2019), participants were asked to predict the average of correlated signals. In the presence (but not absence) of selfish motives, E&K find that participants appeal to the possibility of being confused and making errors that align with their motives, which can aggravate or mitigate correlation neglect. *(iii)* In the *Anchoring* study, based on Enke et al. (2023), participants answered knowledge-based questions facing uninformative anchors. Like in the *Correlation Neglect* study, E&K find evidence for motivated errors: selfish motives lead to an extent of anchoring bias that aligns with subjects appealing to making honest mistakes. In all three studies, the effect sizes of the focal findings are moderate to sizable and statistically significant with $p < 0.01$.

In the course of the *Innsbruck Replication Games*, organized by the *Institute for Replication* (www.i4replication.org), we aimed to evaluate the reliability of E&K's empirical findings in terms of computational reproducibility and robustness to alternative analytical choices in the vein of reproducibility and robustness checks in Brodeur et al. (2024). Using the original data and code, we could regenerate all figures, tables, and results reported in-text in the manuscript and online appendices with full precision. As compared to numerous reports that suggest achieving computational reproducibility poses a significant challenge to the credibility of many empirical studies (Stodden, Seiler, and Ma 2018; Chang and Li 2022; Fišar et al. 2024; Pérignon et al. 2024), the article by E&K stands out favorably. However, limited data availability and a lack of documentation hindered our attempt to assess the robustness of E&K's results as to plausible alternative analytical variations. Our discussion of the challenges encountered can be viewed as a case study of best practices regarding data and code sharing.

2. Computational Reproducibility

There has been considerable debate surrounding the distinct characteristics of different reanalysis concepts (e.g., Patil, Peng, and Leek 2019; Welch 2019). However, the differentiation between reproductions and replications is now widely acknowledged (Dreber and Johannesson 2024). Reproductions strive to regenerate a finding within the same sample using the same methodology, whereas replications seek to reaffirm results in a different sample or through alternative methods (Dreber and Johannesson 2024; Pérignon et al. 2024). Computational reproducibility entails the validation of whether the results obtained from executing the original code on the original data align with the initial reports (National Academies of Sciences, Engineering, and Medicine 2019) and can be viewed as the minimum standard expected for any scholarly published outcome (Christensen and Miguel 2018; Stark 2018).

The assessment of E&K's computational reproducibility commenced with a review of the reproduction package published alongside their article (Exley and Kessler 2024b). The reproduction package comprises five datasets (in *Stata*'s .dta-format), one analysis script (in *Stata*'s .do-format), four *Qualtrics* surveys (.qsf-files) used to collect the data, and a readme file (in .txt-format).

Following E&K's instructions in the readme file, replicators must create subfolders for the output and data and "should expect the code to run within 10 minutes." As it turned out, executing the .do-file was substantially faster: it took less than 20 seconds to run through.[1] No changes to the code were needed, and no additional routines were to be installed manually to make the code run: paths are specified relative to the directory from which the .do-file is opened (i.e., no working directory needs to be set), and required user-written packages are installed on the fly (if not yet installed). No computational errors or versioning conflicts occurred.

The code is complete, i.e., *all* results reported in the manuscript and online appendices are generated by executing the analysis script. All estimates tabulated in E&K are output to publication-ready tables (or table panels) in .tex-format using the user-written routines provided through the "*estout*" package (Jann 2005; 2007), precluding typos in transcribing results to the writeup. Structural comments in E&K's analysis scripts and telling file names of the produced outputs make it

[1] The reproduction was performed in *Stata* (MP18), running under *Windows* 11 (x64) Enterprise, on a *Dell* notebook with an *Intel*® Core™ i7-10610U CPU @1.80GHz (octa-core) and 32GB RAM. As per E&K's readme file, the original analyses were performed in *Stata* (SE 18), running under *macOS Monterey* 12.7.1, on a 2017 *MacBook Air* with an *Intel*® Core™ i5 pro processor @1.8 GHz (dual-core) with 8GB RAM.

straightforward to map the results assembled by executing the .do-file to the results reported in the manuscript and online appendices. In addition, the readme file illustrates the mapping of output files and display items in the manuscript and online appendices.

What it comes down to is that all figures, tables, and in-text results reported in the manuscript and online appendices could be regenerated with full precision using the data and code published alongside the article.[2] Hence, we attest that the results reported in E&K are fully computationally reproducible.

3. Code and Documentation Quality

In the second step of our reproducibility investigation, we carefully reviewed the 2,280 lines of code for potential coding errors, such as variable misspecification or inconsistencies between the code and the descriptions in the manuscript. This exercise, which was supposed to be a breeze, turned out to be more challenging and tedious than expected.

The variable naming conventions in the “raw” data are not self-explanatory, and no codebooks explaining the data are provided.[3] The script file lacks annotations, making it difficult to understand and verify various operations like value assignments, variable manipulations, and if-conditions; structural comments identifying larger code chunks (e.g., "create long data") and display items (e.g., "figure 1") are the only exceptions. The code involves multiple copied-and-pasted operations, which makes it generally prone to error, and the coding style seems maverick and somewhat inefficient. For example, the authors define relative directory paths as local macros, which prevents running chunks of the code without executing the macro definitions at the very top of the .do-file.

After all, despite the pitfalls discussed above, we did not find any obvious errors in the coding. However, it's important to note that our assessment is limited by the lack of annotations in the analysis code and the absence of codebooks, which prevents us from confidently declaring that the analyses are flawless and reliable.

[2] Since the results obtained from the reproduction are identical for each and every of the reports in E&K’s manuscript and online appendices, we—in the interest of conciseness—abstain from echoing them in this report.

[3] Notably, the replication package comprises the *Qualtrics* surveys (.qsf files) that were utilized to gather the data. Examining these files would likely provide insights into variable definitions. However, collating the surveys with the data and code files would be cumbersome and time-consuming, and would require a subscription to a paid *Qualtrics* plan.

4. Data Availability

The five datasets in the replication kit are all named with the suffix "_raw." Inspecting the data, however, makes clear that the data files are *not* raw but have been preprocessed: *(i)* the data are only available in .dta-format,[4] *(ii)* custom variable labels are attached to some (but not all) variables,[5] and *(iii)* data that is recorded in *Qualtrics* by default (e.g., response identifiers, timestamps, status indicators) are lacking from the datasets.

A lack of authentic raw data opens the door to potential concerns about the integrity and originality of the publicly shared data records. Without access to the raw data files (and the code used to preprocess them), it is impossible to rule out that data have been modified, falsified, or even fabricated (Fanelli 2009). We refrain from accusing E&K of any questionable research practices or misconduct but emphasize that the absence of unprocessed raw data necessitates further investigation into data integrity to ensure the study's results are reliable.

Notably, all observations in the data are complete.[6] In light of various reports of non-negligible attrition in studies using crowdsourced convenience samples (see, e.g., Crump, McDonnell, and Gureckis 2013; Albert and Smilek 2023), a dropout rate of 0% appears improbable; incomplete data records were likely dropped from the data. Unattended variation in attrition rates across conditions jeopardizes indirect control, with potentially severe consequences for inference (Zhou and Fishbach 2016). Randomization checks could rule out these concerns but are precluded due to incomplete data. Relatedly, in light of concerns about impaired data quality of crowd-sourced samples due to inattentive participants and bots (Zhou and Fishbach 2016; Chmielewski and Kucker 2020; Aguinis, Villamor, and Ramani 2021; Peer et al. 2022; Webb and Tangney 2022; Douglas, Ewell, and Brauer 2023), it is lamentable that there are no records that allow data quality to be assessed (e.g., HIT approval rates, attention checks, IP quality checks, response times, etc.). Based on the available data, it is virtually impossible to affirm data quality.

---

[4] While the list of data export formats supported by *Qualtrics* is fairly long—featuring several open file formats (e.g., .csv, .tsv, .xml, or .json) and proprietary formats such as .xlsx (*Microsoft Excel*) or .sav (*SPSS*)—, to the best of our knowledge, *Qualtrics* does not allow for .dta exports.

[5] For some variables (e.g., treatment and control condition indicators), the labels are defined in *LaTeX* syntax to be parsed in the automated output of results in .tex-format. Since, e.g., condition labels have changed relative to previous versions of the manuscript (see Section 2.3 for details), it becomes obvious that the labels were attached ex post.

[6] E&K are silent about dropouts. Only in the notes to Table A.1 in the online appendix, they mention that "31 prior subjects who participated" were excluded from the data collection for the *Adding Study* in December 2019 "due to a recruitment error." The excluded observations do not show up in the "raw" data.

Furthermore, relevant participant-level data is missing in the data. *Qualtrics* records important information such as response identifiers, starting and ending timestamps, and status indicators by default. However, these records are not included in the datasets, ruling out data integrity checks. For example, the sorting of the dataset cannot be straightforwardly reproduced based on any combination of variables in the "*Adding*" dataset: While treatment conditions appear to be randomly assigned for the first half, conditions show up in sequence further down the dataset. Although the block-wise listing of treatments can be rationalized by E&K's summary of the data collection in the table notes to Table A.1 in the online appendix, there is no means to verify that the data is sorted in the order in which it was collected, is complete, and has not been modified.

5. Evolution of Exley and Kessler's (2024) Manuscript

In Section E of their online appendix, E&K document the evolution of their article and transparently highlight that the manuscript has substantially changed across several previous draft versions. Specifically, they disclose that—since the first submission to the *American Economic Review* in 2017—several studies were added and removed under the guidance of the handling editor and reviewers throughout the publication journey. The authors deserve commendation for this exemplary level of transparency (Miguel et al. 2014; Aczel et al. 2020), particularly in light of the lack of preregistration.[7] The four previous working paper versions of the article (Exley and Kessler 2017; 2018; 2019; 2022) are referenced,[8] with links provided, and are still accessible when writing this report.

The most focal changes (i.e., added/dropped studies/treatments) from previous to subsequent versions are summarized in Section E of E&K's online appendix (pp. 48–49). We collated the results reported in the published version with those reported in earlier drafts and vetted E&K's summary of the paper's evolution for consistency and completeness. While the labels for treatments and the way information is presented have slightly changed across the different versions of the paper, the sample sizes for each treatment group and the analytical methods used have stayed the same. As a result, the empirical estimates remained unchanged

[7] Given that the initial experiments took place in 2017, or possibly even earlier, a time when preregistration and pre-analysis plans were not yet a common practice, it comes as no surprise that the studies were not preregistered.

[8] The first working paper version (Exley and Kessler 2017) was circulated under the title "*The Better is the Enemy of the Good*;" all subsequent versions of the paper (Exley and Kessler 2018; 2019; 2022; 2024b) were circulated under the title "*Motivated Errors*."

between the previous and subsequent versions of the manuscript. The consistency in the results reported across the different working paper versions, along with the high level of transparency, alleviates concerns about selective reporting, publication bias, and the "chrysalis effect" (see, e.g., John, Loewenstein, and Prelec 2012; Simonsohn, Nelson, and Simmons 2014; O'Boyle, Banks, and Gonzalez-Mulé 2017) and enhances the credibility of the empirical results.

6. Robustness Reproduction

The results of recent crowd-science projects suggest that analytical heterogeneity can be substantial (e.g., Silberzahn et al. 2018; Botvinik-Nezer et al. 2020; Huntington-Klein et al. 2021; Breznau et al. 2022; Menkveld et al. 2024), adding a layer of uncertainty potentially undermining the generalizability of empirical research findings (Kenny and Judd 2019; Yarkoni 2020; Holzmeister et al. 2024a). In light of the discussions surrounding heterogeneity, its opportunistic misuse, and—more generally—the credibility of empirical scholarly claims (see, e.g., Simmons, Nelson, and Simonsohn 2011; Wicherts et al. 2016; Brodeur, Cook, and Heyes 2020; 2022), we aimed to vet the robustness of E&K's empirical results as to alternative analytical specifications.

While reviewing the original data and code, we soon learned that the lack of annotation of the analysis scripts and the unavailability of relevant data (as discussed in sections 3 and 4) would add serious limitations to our goal. Worthwhile integrity checks and robustness analyses are precluded due to the unavailability of relevant data; randomization and balance checks cannot be performed due to missing data on incomplete observations; the lack of demographic data rules out subsample analyses that could shed light on the potential impact of moderators; missing records on data quality indicators preclude sensitivity analyses excluding likely unattentive participants; etc.

In a nutshell, the data published alongside the article does not allow for conducting *meaningful* robustness analyses other than those already reported in the article. Since E&K's experimental designs implicate clear identification strategies, the degrees of freedom in the statistical analysis are held at bay, and the sensible analytical space is comparatively small.

Furthermore, it should also be noted that the three studies were decently powered, even if the article is silent about how the sample sizes were determined.[9] For the focal hypothesis tests,[10] 45.4% (*Adding*), 109.6% and 23.3% (*Correlation Neglect*), and 78.5% and 48.3% (*Anchoring*) of the actual sample sizes would have been needed to detect effect sizes as large as the ones reported in the article with 90% statistical power at a two-tailed 5% significance level.[11] Hence, the primary tests in E&K (except the test of "exacerbated bias" in the *Correlation Neglect* study) were highly powered ($\pi = 0.90$) to detect smaller effects than the ones reported.[12] With the $|t|$-statistics of the focal tests ranging from 3.10 to 6.71, there is a decent "buffer" for coefficient estimates or standard errors to vary across alternative analytical models without impacting the conclusions. Thus, in light of the tight identification strategy and high statistical power, there is no point in carrying out robustness tests the data would allow for; the claims can obviously be expected to be robust to alternative econometric specifications (e.g., a logit model instead of a linear probability model, HC2 or HC3 errors instead of HC1 errors, etc.).

[9] The main results of the *Adding* study were estimated based on a sample of $n = 397$ participants ($n = 199$ in the "Charity/Charity" condition and $n = 198$ in the "Self/Charity" condition), who completed 36 trials each. The primary results of the *Correlation Neglect* study, were estimated based on a sample of 1,200 participants randomly assigned to one of three conditions, answering ten questions each ($n = 408$ in the "Control" condition, $n = 396$ in the "Exacerbate Bias" condition, and $n = 396$ in the "Mitigate Bias" condition). The ample in the *Anchoring* study comprised 1,195 participants answering four questions each ($n = 398$ in the "Control" condition, $n = 397$ in the "Exacerbate Bias" condition, and $n = 400$ in the "Mitigate Bias" condition).

[10] Although the estimates are only reported in the online appendix, we consider the estimates of differences in the effect size of the stimuli between conditions with and without selfish motives (i.e., the interaction terms of stimuli and the treatment manipulations) the focal hypotheses tests concerning the main conclusions drawn.

[11] The estimates are based on the calculations used in Holzmeister et al. (2024b); see Szucs and Ioannidis (2017) for the underlying derivations. The sample size and minimal detectable effect size calculation refer to the following estimates in the article: *(i) Adding* study: coefficient estimate (standard error) of "Charity/Charity × (+0)" in Table B.1, reported as 0.07 (0.02); *(ii) Correlation Neglect* study: estimates of "Low E1 × Exacerbate Bias" and "Low E1 × Mitigate Bias" in Table B.5, reported as −5.30 (1.71) and 11.80 (1.76), respectively; and *(iii) Anchoring* study: estimates of "Low Anchor × Exacerbate Bias" and "Low Anchor × Mitigate Bias" in Table B.6, reported as −7.98 (2.18) and 10.82 (2.32), respectively. To obtain precise estimates, we used non-rounded figures obtained from regenerating E&K's results based on their data and code instead of the estimates reported in the manuscript (which are rounded to two decimal places).

[12] Based on the reported standard errors, the post-hoc minimum detectable effect sizes (assuming 90% power at the two-tailed 5%-level) are 5.0 *pp* (67.4% of the reported effect size in the *Adding* study), 5.5 and 5.7 (104.7% and 48.3% of the effect sizes in the *Correlation Neglect* study), and 7.1 and 7.5 (88.6% and 69.5% of the effect sizes in the *Anchoring* study).

7. Summary and Conclusion

The empirical results reported in E&K's article (and the accompanying online appendices) were fully computationally reproducible using the data and code made publicly available in the *American Economic Review*'s data repository. In other words, all estimates and display items generated using the replication kit match the results reported in the publication with full precision. Although considered a minimum standard in scholarly publishing, empirical investigations suggest that computational reproducibility is all but straightforward to achieve, making E&K's article rank favorably in comparison within and across fields (see, e.g., Trisovic et al. 2022; Fišar et al. 2024).

Yet, the data and code shared alongside the publication (Exley and Kessler 2024b) come with several limitations. The variables comprised in the dataset are neither named in a self-explanatory way nor conveyed in codebooks, the analysis script is barely annotated, making it challenging to follow the authors' coding, and various relevant participant-level records are missing from the data files. After all, these limitations imply that the data and code shared alongside the article are barely reusable for any investigation other than computational reproducibility checks and thwarted our goal to vet the empirical findings' analytical robustness. Our discussions of the caveats in examining the reliability of E&K's empirical claims are intended to shed some light on best practices regarding the do's and don'ts in sharing data and code (also see Pérginon et al. 2024). Nonetheless, the fact that E&K demonstrate "motivated errors" in both simple and more intricate settings, their laudable transparency regarding the article's evolution, the tight link between the econometric models and experimental designs, and the tests' adequate statistical power help to overcome potential concerns about the robustness of the estimates and lend credibility to the empirical claims.

References

Aczel, B., B. Szaszi, A. Sarafoglou, Z. Kekecs, Šimon Kucharský, D. Benjamin, C. D. Chambers, et al. 2020. "A Consensus-Based Transparency Checklist." *Nature Human Behaviour* 4 (1): 4–6. https://doi.org/10/ggd32c.

Aguinis, H., I. Villamor, and R. S. Ramani. 2021. "MTurk Research: Review and Recommendations." *Journal of Management* 47 (4): 823–37. https://doi.org/10/ghkskx.

Albert, D. A., and D. Smilek. 2023. "Comparing Attentional Disengagement between Prolific and MTurk Samples." *Scientific Reports* 13 (1): 20574. https://doi.org/10/gt6p7c.

Botvinik-Nezer, R., F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, et al. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature* 582 (7810): 84–88. https://doi.org/10/ggwrvt.

Breznau, N., E. M. Rinke, A. Wuttke, H. H. V. Nguyen, M. Adem, J. Adriaans, A. Alvarez-Benjumea, et al. 2022. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty." *Proceedings of the National Academy of Sciences* 119 (44): e2203150119. https://doi.org/10/gq5vs5.

Brodeur, A., N. Cook, and A. Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60. https://doi.org/10/ghg83w.

———. 2022. "We Need to Talk about Mechanical Turk: What 22,989 Hypothesis Tests Tell Us about Publication Bias and p-Hacking in Online Experiments." Working Paper. https://doi.org/10/nd9h.

Brodeur, A., D. Mikola, N. Cook, and et al. 2024. "Mass Reproducibility and Replicability: A New Hope." Working Paper. https://doi.org/10/nd9j.

Chang, A. C., and P. Li. 2022. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Often Not.'" *Critical Finance Review* 11 (1): 185–206. https://doi.org/10/gsqfkp.

Chmielewski, M., and S. C. Kucker. 2020. "An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results." *Social Psychological and Personality Science* 11 (4): 464–73. https://doi.org/10/gf92b6.

Christensen, G., and E. Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–80. https://doi.org/10/gfbkkh.

Crump, M. J. C., J. V. McDonnell, and T. M. Gureckis. 2013. "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research." *PLoS One* 8 (3): e57410. https://doi.org/10/f4qw94.

Douglas, B. D., P. J. Ewell, and M. Brauer. 2023. "Data Quality in Online Human-Subjects Research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA." *PLoS One* 18 (3): e0279720. https://doi.org/10/grx5s2.

Dreber, A., and M. Johannesson. 2024. "A Framework for Evaluating Reproducibility and Replicability in Economics." *Economic Inquiry* online first. https://doi.org/10/gt3vmw.

Enke, B., U. Gneezy, B. Hall, D. Martin, V. Nelidov, T. Offerman, and Jeroen van de Ven. 2023. "Cognitive Biases: Mistakes or Missing Stakes?" *The Review of Economics and Statistics* 105 (4): 818–32. https://doi.org/10/gt7zds.

Enke, B., and F. Zimmermann. 2019. "Correlation Neglect in Belief Formation." *The Review of Economic Studies* 86 (1): 313–32. https://doi.org/10/gf7rmt.

Exley, C. L., and J. B. Kessler. 2017. "The Better Is the Enemy of the Good." Working Paper. https://stanford.io/3MjvLlr.

———. 2018. "Motivated Errors." Working Paper. https://stanford.io/4cpx2lk.

———. 2019. "Motivated Errors." Working Paper. https://bit.ly/4fGLt7B.

———. 2022. "Motivated Errors." Working Paper. https://bit.ly/4dDTCYj.

———. 2024a. "Motivated Errors." *American Economic Review* 114 (4): 961–87. https://doi.org/10/gt2jb6.

———. 2024b. "Replication Package: Data and Code for 'Motivated Errors.'" Inter-University Consortium for Political and Social Research (ICPSR). https://doi.org/10/m4wz.

Fanelli, D. 2009. "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data." *PLoS One* 4 (5): e5738. https://doi.org/10/bn5pnj.

Fišar, M., B. Greiner, C. Huber, E. Katok, A. I. Ozkes, and the Management Science Reproducibility Collaboration. 2024. "Reproducibility in Management Science." *Management Science* 70 (3): 1343–2022. https://doi.org/10/gtbtwh.

Holzmeister, F., M. Johannesson, R. Böhm, A. Dreber, J. Huber, and M. Kirchler. 2024a. "Heterogeneity in Effect Size Estimates." *Proceedings of the National Academy of Sciences* 121 (32): e2403490121. https://doi.org/10/gt5nkn.

Holzmeister, F., M. Johannesson, C. F. Camerer, Y. Chen, T.-H. Ho, S. Hoogeveen, J. Huber, et al. 2024b. "Replication by Plebiscite: Examining the Replicability of Online Experiments Selected by a Decision Market." *Mimeo*.

Huntington-Klein, N., A. Arenas, E. Beam, M. Bertoni, J. R. Bloem, P. Burli, N. Chen, et al. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59 (3): 944–60. https://doi.org/10/gk39mh.

Jann, B. 2005. "Making Regression Tables from Stored Estimates." *The Stata Journal* 5 (3): 288–308. https://doi.org/10/gf3dcc.

———. 2007. "Making Regression Tables Simplified." *The Stata Journal* 7 (2): 227–44. https://doi.org/10/gf3dcb.

John, L. K., G. Loewenstein, and D. Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–32. https://doi.org/10/f33h6z.

Kenny, D A., and C M. Judd. 2019. "The Unappreciated Heterogeneity of Effect Sizes: Implications for Power, Precision, Planning of Research, and Replication." *Psychological Methods* 24 (5): 578–89. https://doi.org/10/gf8936.

Menkveld, A. J., A. Dreber, F. Holzmeister, J. Huber, M. Johannesson, M. Kirchler, M. Razen,

et al. 2024. “Non-Standard Errors.” *Journal of Finance* 79 (3): 2239–2390. https://doi.org/10/gtrm5b.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, et al. 2014. “Promoting Transparency in Social Science Research.” *Science* 343 (6166): 30–31. https://doi.org/10/gdrcpz.

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, D.C.: National Academies Press. https://doi.org/10/c5jp.

O’Boyle, E. H., G. C. Banks, and E. Gonzalez-Mulé. 2017. “The Chrysalis Effect: How Ugly Initial Results Metamorphosize into Beautiful Articles.” *Journal of Management* 43 (2): 376–99. https://doi.org/10/gf2k59.

O’Grady, C. 2024. “Embattled Harvard Honesty Professor Accused of Plagiarism.” *Science (ScienceInsider)*. https://doi.org/10/gt6tqz.

Patil, P., R. D. Peng, and J. T. Leek. 2019. “A Visual Tool for Defining Reproducibility and Replicability.” *Nature Human Behaviour* 3 (7): 650–52. https://doi.org/10/ggrnj8.

Peer, E., D. Rothschild, A. Gordon, Z. Evernden, and E. Damer. 2022. “Data Quality of Platforms and Panels for Online Behavioral Research.” *Behavior Research Methods* 54 (4): 1643–62. https://doi.org/10/gn3m43.

Pérignon, C., O. Akmansoy, C. Hurlin, A. Dreber, F. Holzmeister, J. Huber, M. Johannesson, et al. 2024. “Computational Reproducibility in Finance: Evidence from 1,000 Tests.” *The Review of Financial Studies*, hhae029. https://doi.org/10/gt4jg7.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, et al. 2018. “Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results.” *Advances in Methods and Practices in Psychological Science* 1 (3): 337–56. https://doi.org/10/gd2429.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–66. https://doi.org/10/bxbw3c.

Simonsohn, U., L. D. Nelson, and J. P. Simmons. 2014. “P-Curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143 (2): 534–47. https://doi.org/10/gffnn9.

Stark, P. B. 2018. “Before Reproducibility Must Come Preproducibility.” *Nature* 557 (7707): 613–613. https://doi.org/10/gdh9xc.

Stodden, V., J. Seiler, and Z. Ma. 2018. “An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility.” *Proceedings of the National Academy of Sciences* 115 (11): 2584–89. https://doi.org/10/gc8gkw.

Szucs, D., and J. P. A. Ioannidis. 2017. “Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature.” *PLoS Biology* 15 (3): e2000797. https://doi.org/10/b4r4.

Trisovic, A., M. K. Lau, T. Pasquier, and M. Crosas. 2022. “A Large-Scale Study on Research Code Quality and Execution.” *Scientific Data* 9 (1): 60. https://doi.org/10/grhw73.

Webb, M. A., and J. P. Tangney. 2022. “Too Good to Be True: Bots and Bad Data from Mechanical Turk.” *Perspectives on Psychological Science*, 174569162211200. https://doi.org/10/gq7nw4.

Welch, I. 2019. “Reproducing, Extending, Updating, Replicating, Reexamining, and Reconciling.” *Critical Finance Review* 8 (1–2): 301–4. https://doi.org/10/gttsh4.

Wicherts, J. M., C. L. S. Veldkamp, H. E. M. Augusteijn, M. Bakker, R. C. M. van Aert, and M. A. L. M. van Assen. 2016. “Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking.” *Frontiers in Psychology* 7. https://doi.org/10/gc5sjn.

Yarkoni, T. 2020. “The Generalizability Crisis.” *The Behavioral and Brain Sciences* 45:e1. https://doi.org/10/gh28p8.

Zhou, H., and A. Fishbach. 2016. “The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (yet False) Research Conclusions.” *Journal of Personality and Social Psychology* 111 (4): 493–504. https://doi.org/10/f854b9.